THERMOV: THERMAL-ONLY OPEN-VOCABULARY SEMANTIC SEGMENTATION VIA UNSUPERVISED CROSS-DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary semantic segmentation (OVSS) enables pixel-level recognition of arbitrary visual concepts specified by text. However, existing OVSS methods are predominantly developed for RGB imagery, while prior studies involving thermal data typically rely on co-registered RGB-thermal inputs. As a result, open-vocabulary segmentation from thermal observations alone remains largely unexplored. To bridge this gap, we propose ThermOV, a two-stage unsupervised cross-domain adaptation framework that transfers the open-vocabulary segmentation capability of the RGB-trained foundation model SAM3 to the thermal domain without requiring pixel-level annotations of real thermal images. In the first stage, labeled RGB images are used to generate synthetic thermal images to provide an initial adaptation to thermal appearance. In the second stage, the model is further adapted using unlabeled real thermal images within an EMA teacher-student framework. To preserve open-vocabulary generalization during this process, we construct a diverse prompt bank that excludes categories overlapping with the target evaluation classes and jointly optimize pseudo-label supervision, semantic prompt consistency and teacher-student prediction consistency. Experiments on FMB and TIR-SS show that ThermOV improves over SAM3 by 5.97 and 9.30 mIoU, respectively, and outperforms the strongest RGB OVSS baseline adapted using labeled real thermal data by 3.81 and 0.54 mIoU. These results demonstrate the effectiveness of ThermOV for thermal-only OVSS.

1 INTRODUCTION

Thermal infrared imaging provides valuable visual information in challenging environments. Unlike RGB sensors, which rely on visible illumination, thermal cameras capture radiation emitted by objects and can operate under conditions such as darkness, glare, and smoke. Thermal imagery is therefore widely used in safety-critical applications, including autonomous driving, nighttime surveillance, and robotic perception Son et al. (2026); Zhu et al. (2025); Zhang et al. (2025).

Generally, the abovementioned applications require pixel-level scene understanding across diverse semantic categories encountered during deployment. However, conventional semantic segmentation models are restricted to a predefined set of categories, limiting their ability to accommodate evolving recognition needs. Open-vocabulary semantic segmentation (OVSS) addresses this limitation by extending pixel-level recognition beyond a fixed vocabulary through text-specified categories. Recent advances in OVSS have primarily focused on RGB imagery (Liang et al., 2023; Xu et al., 2023; Cho et al., 2024) and largely rely on pretrained vision-language models (VLMs) such as CLIP (Radford et al., 2021), whose aligned visual and linguistic representations support generalization beyond the categories annotated during segmentation training.

Existing attempts to extend OVSS to thermal imagery mainly adopt the RGB-T setting and require aligned RGB and thermal inputs (Zhao et al., 2024; Yu et al., 2025). In contrast, OVSS from thermal observations alone remains largely underexplored. As illustrated in Fig. 1, the RGB-to-thermal shift in appearance and feature distributions (Chen et al., 2022; Gan et al., 2023) can disrupt the vision-language alignment learned from RGB data, hindering the direct transfer of RGB-trained OVSS models to the thermal domain and resulting in missed regions and category confusion.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

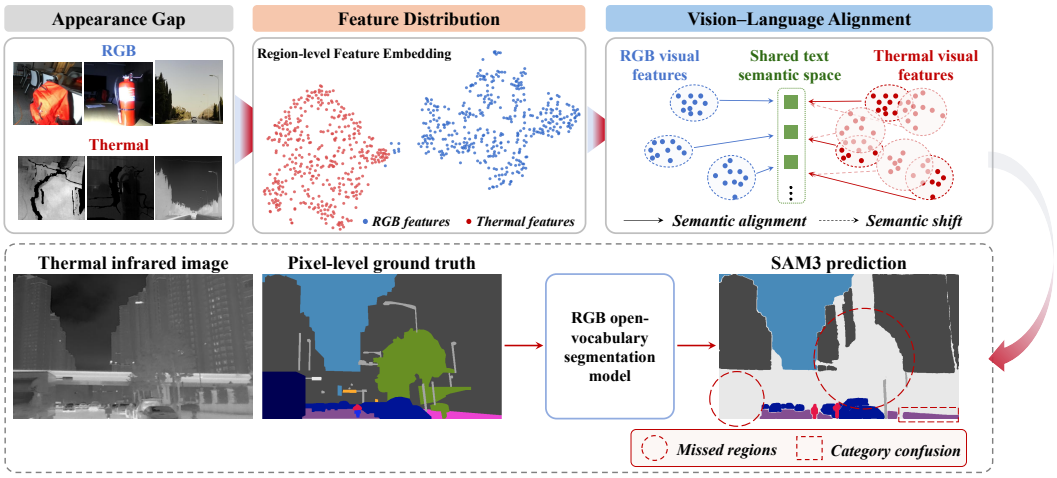

Figure 1: Conceptual illustration of the RGB-to-thermal transfer gap, where appearance, feature, and vision-language discrepancies hinder direct transfer of open-vocabulary knowledge.

A straightforward approach to thermal-only OVSS is to adapt an RGB-trained OVSS model using pixel-level annotations of thermal images. However, unlabeled thermal imagery is readily available in practice, whereas pixel-level annotations remain scarce because obtaining them is costly and time-consuming Chen et al. (2022); Gan et al. (2023). Thus this raises a key question: **How can abundant unlabeled thermal images be effectively exploited to enable thermal-only OVSS?** To address this challenge, we propose ThermOV, a two-stage unsupervised cross-domain adaptation framework that transfers the open-vocabulary segmentation capability of an RGB-trained model to the thermal domain without requiring pixel-level annotations of real thermal images. In the first stage, ThermOV performs an initial adaptation to thermal appearance using synthetic thermal images generated from labeled RGB data. Building on this initialization, the second stage exploits unlabeled real thermal images to further adapt the model to the real thermal domain while preserving its open-vocabulary recognition capability.

Our main contributions are summarized as follows:

- We present, to the best of our knowledge, the first systematic study of unsupervised thermal-only OVSS adaptation. We further establish a unified evaluation benchmark suite on FMB (Liu et al., 2023) and TIR-SS (Chen et al., 2022), comprising 16 semantic concepts and 48 diverse prompts.
- We propose ThermOV, a two-stage cross-domain adaptation framework for thermal-only OVSS. It first learns from labeled pseudo-thermal images and then adapts to unlabeled real thermal images, eliminating the need for real thermal pixel annotations while preserving open-vocabulary recognition.
- Extensive experiments demonstrate that ThermOV improves over SAM3 by 5.97 and 9.30 mIoU on FMB and TIR-SS, respectively, and outperforms the strongest RGB OVSS baselines adapted using labeled real thermal data by 3.81 and 0.54 mIoU.

2 Related Work

Open-Vocabulary Semantic Segmentation. Existing OVSS methods can be broadly categorized into training-based and training-free approaches. Training-based methods use segmentation supervision to fine-tune VLMs or train task-specific modules, such as decoders and adapters, to transform image-level vision-language representations into dense pixel predictions. Representative methods include OVSeg (Liang et al., 2023), SAN (Xu et al., 2023), CAT-Seg (Cho et al., 2024), and MaskAdapter (Li et al., 2025). In contrast, training-free methods keep pretrained model parameters frozen and generate segmentation predictions by modifying attention operations, combining complementary foundation-model features, or refining category predictions at inference time. For example, NA-

CLIP (Hajimiri et al., 2025) introduces neighborhood-aware attention into CLIP to improve dense localization, Trident (Shi et al., 2025) combines representations from CLIP, DINO, and SAM to address limited spatial resolution, and FreeCP (Chen et al., 2025) mitigates class redundancy and vision-language ambiguity through class purification.

More recently, promptable segmentation foundation models have extended open-vocabulary recognition to concept-level mask prediction. SAM3 (Carion et al., 2026) introduces Promptable Concept Segmentation, which detects and segments all instances matching a concept specified by a short noun phrase, an image exemplar, or both. We adopt SAM3 as the base model of ThermOV because it directly unifies text-guided concept recognition and mask prediction. Nevertheless, SAM3 and most existing OVSS methods are developed primarily for RGB imagery, and their open-vocabulary capabilities do not transfer reliably to the thermal domain.

Open-Vocabulary Perception with Thermal Imagery. Existing studies on open-vocabulary thermal perception have primarily focused on image-level representation learning and region-level object detection. For example, Thermal-Det (Ranasinghe et al., 2026) transfers open-vocabulary detection knowledge to thermal imagery using synthetic thermal data and cross-modal distillation, while T-CLIP (Qazi et al., 2026) adapts CLIP representations to improve thermal image-text alignment.

At the pixel level, existing open-vocabulary segmentation methods involving thermal imagery remain confined to the RGB-T multimodal setting. These methods require paired and spatially aligned RGB and thermal inputs and exploit their complementary information to improve segmentation performance. Representative approaches include OpenRSS (Zhao et al., 2024) and Open-RGBT (Yu et al., 2025), both of which require RGB-T inputs at inference time. AetherVision-Bench (Sikdar et al., 2025) evaluates existing OVSS models across RGB and infrared datasets and reveals substantial performance degradation under sensor and viewpoint shifts. However, it serves primarily as an evaluation benchmark rather than a method for adapting OVSS models to thermal imagery. Consequently, thermal-only OVSS, where only thermal images are required at inference time, has yet to be systematically studied.

Unsupervised RGB-Infrared Adaptation. Unsupervised domain adaptation (UDA) aims to transfer models from labeled source domains to unlabeled target domains and has been extensively studied for semantic segmentation. Existing approaches, such as DACS (Tranheden et al., 2021) and DAFormer (Hoyer et al., 2022), mainly address closed-set domain shifts through self-training and domain-specific optimization. For RGB-to-thermal transfer, Gan et al. (Gan et al., 2023) investigate thermal semantic segmentation without thermal annotations or paired RGB-T data. However, these methods are mainly designed for fixed-category recognition and do not consider open-vocabulary vision-language scenarios.

Recent studies further extend domain adaptation to open-vocabulary segmentation, including SemLA (Qorbani et al., 2025) and VocAlign (Mazzucco et al., 2025). These methods mainly focus on distribution shifts within the same visual modality and provide limited consideration of the larger cross-modal gap between RGB and thermal imagery. Therefore, thermal-only open-vocabulary semantic segmentation with real unlabeled thermal images remains largely unexplored.

3 METHOD

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Problem Formulation. We study thermal-only OVSS under an unsupervised target-domain setting. Let the labeled RGB source domain be $\mathcal{D}_s = \{(\mathbf{x}_i^R, \mathbf{y}_i^R)\}_{i=1}^{N_s}$, where $\mathbf{x}_i^R$ and $\mathbf{y}_i^R$ denote an RGB image and its pixel-level semantic annotation, respectively. The real thermal target domain is $\mathcal{D}_t = \{\mathbf{x}_j^T\}_{j=1}^{N_t}$, where no pixel-level annotations are available during training. Given a text-defined category vocabulary $\mathcal{C} = \{c_k\}_{k=1}^{K}$, our goal is to learn a segmentation model F_θ that predicts the semantic concept of each pixel in a thermal image, i.e., $\hat{\mathbf{y}}^T = F_\theta(\mathbf{x}^T, \mathcal{C})$. The key challenge is to bridge the RGB-to-thermal domain gap without target-domain annotations while preserving the open-vocabulary recognition capability learned from RGB data.

Framework Overview. As illustrated in Fig. 2, ThermOV initializes from SAM3 (Carion et al., 2026) and performs progressive RGB-to-thermal adaptation through two stages. The first stage

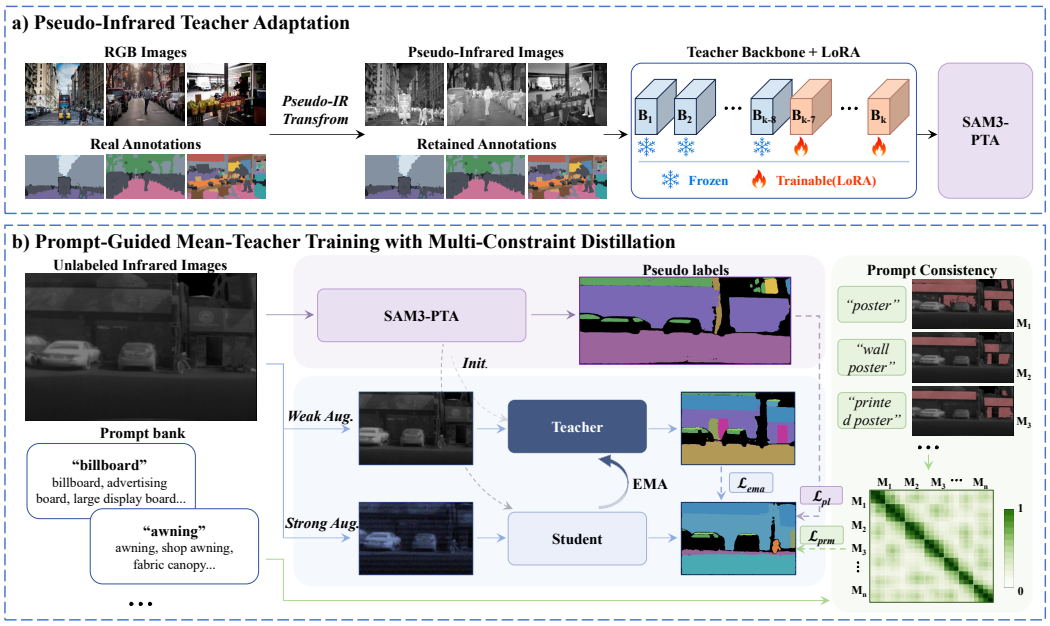

Figure 2: Two-stage unsupervised adaptation pipeline of ThermOV for thermal-only OVSS.

performs pseudo-thermal supervised initialization, establishing thermal-domain visual representations from labeled RGB data. The second stage further exploits unlabeled real thermal images for adaptation, jointly optimizing pseudo-label supervision, teacher-student prediction consistency, and semantic prompt consistency across different expressions of the same concept.

3.2 Stage I: Pseudo-Thermal Domain Initialization

SAM3 is originally trained on large-scale RGB imagery, and directly adapting its visual representation to thermal images is challenging due to the substantial appearance discrepancy between the two modalities. To alleviate this cross-modal discrepancy, we first construct pseudo-thermal supervision from labeled RGB images in COCO-Stuff (Caesar et al., 2018) and perform parameter-efficient adaptation on the visual representation of SAM3.

Pseudo-Thermal Data Generation. Each training sample from COCO-Stuff is represented as $(\mathbf{x}^R, p, \mathbf{m})$, where $\mathbf{x}^R$ denotes the RGB image, p denotes the semantic prompt, and $\mathbf{m}$ denotes the corresponding pixel-level mask. We first convert the RGB image into a single-channel luminance representation using the standard luminance conversion, $\mathbf{Y} = 0.299R + 0.587G + 0.114B$. This conversion removes color information while preserving the spatial structure of the original scene.

Subsequently, we construct a category-aware thermal prior map $\mathbf{\Delta}$ based on the semantic labels, with the category-specific prior values provided in *Suppl.* Table S1. The prior adjusts the relative intensity responses among different semantic regions to simulate the temperature-related appearance variations in thermal imagery. The thermal-enhanced representation is formulated as:

$$\mathbf{Y}_t = \text{clip}(\mathbf{Y} + \mathbf{\Delta}, 0, 1). \tag{1}$$

Image-level appearance perturbations are then applied to $\mathbf{Y}_t$ to simulate variations in thermal imaging conditions, with implementation details provided in *Suppl.* Table S2. Finally, the resulting single-channel thermal-style image is replicated across three channels to obtain the pseudo-thermal image $\tilde{\mathbf{x}}^T$. An example is shown in Fig. 3. Through the above transformation, pseudo-thermal supervision is represented as $(\tilde{\mathbf{x}}^T, p, m)$.

Parameter-Efficient Visual Adaptation. Using the constructed pseudo-thermal supervision, we adapt SAM3's visual encoder to thermal imagery in a parameter-efficient manner. Specifically, we freeze all original SAM3 parameters and insert LoRA adapters (Hu et al., 2022) into the query, key, and value projection layers of the last 12 Transformer blocks in the visual encoder. Only the LoRA



(a) RGB Input (b) Pseudo-Thermal Image (c) Pixel-level Annotation

Figure 3: Examples of pseudo-thermal images generated from labeled RGB data.

parameters are updated during training, allowing the model to capture thermal appearance characteristics while retaining its pretrained open-vocabulary knowledge. The resulting model, denoted as SAM3-PTA, serves as the initialization for subsequent adaptation to real thermal images.

3.3 Stage II:Unlabeled Thermal Domain Adaptation

Prompt-Guided Pseudo-Label Generation. Given an unlabeled real thermal image $\mathbf{x}^T$, we first introduce a prompt-guided strategy to generate pseudo-labels for subsequent training. Specifically, we manually curate 76 semantic concepts covering common objects in urban environments, road scenes, infrastructure, and building components. For each semantic concept c_i, we use a large language model to generate three prompt variants, forming a semantic prompt bank $\mathcal{P} = \{p_{ij} \mid i = 1, \ldots, 76, \ j = 1, 2, 3\}$, with representative examples provided in *Suppl.* Table S4. Then SAM3-PTA predicts a corresponding segmentation mask for each semantic concept c_i:

$$\hat{\mathbf{m}}_i = \text{SAM3-PTA}(\mathbf{x}^T, c_i) \tag{2}$$

Candidate masks that pass confidence filtering are merged into a binary pseudo mask $\hat{\mathbf{m}}_i$, which is shared across all prompt variants associated with the same semantic concept c_i. The resulting triplet $(\mathbf{x}^T, c_i, \hat{\mathbf{m}}_i)$ provides supervision for subsequent adaptation to real thermal images, while the prompt variants are further used to construct the semantic prompt consistency constraint described in Sec. 3.4.

EMA Teacher-Student Self-Training. Using the generated pseudo-labels, we adapt SAM3-PTA to the real thermal domain under an unsupervised domain adaptation setting. To stabilize this adaptation process, we employ an EMA teacher-student framework (Tarvainen & Valpola, 2017). Both models are initialized with the parameters of SAM3-PTA. During training, the student parameters are updated through backpropagation, while the teacher parameters are updated as the exponential moving average of the student parameters:

$$\theta_T^{(t)} = \alpha\theta_T^{(t-1)} + (1-\alpha)\theta_S^{(t)}, \tag{3}$$

where $\theta_T^{(t)}$ and $\theta_S^{(t)}$ denote the teacher and student parameters at training step t, respectively, and α is the momentum coefficient. This EMA update smooths fluctuations in the student parameters, enabling the teacher to provide more stable prediction targets during adaptation.

Following standard practice in teacher-student self-training (Zhao et al., 2023), the teacher receives a weakly augmented thermal view, while the student receives a slightly stronger view with random horizontal flipping and mild photometric perturbations.Detailed augmentation settings are provided in *Suppl.* Table S6. This stage is optimized using the training objectives described in Sec. 3.4.

3.4 Multi-Constraint Joint Training Strategy

During Stage II, the student model is optimized using three complementary objectives. Pseudo-label supervision provides direct target-domain semantic guidance, EMA teacher–student consistency encourages stable predictions under image perturbations, and semantic prompt consistency improves robustness to diverse textual expressions.

Pseudo-Label Supervision. The pseudo-labels generated by Eq. 2 provide direct pixel-level supervision for target-domain adaptation. Since only the student model is optimized through backpropagation, we first define its pseudo-label supervision loss. Given a training triplet $(\mathbf{x}^T, p_i, \hat{\mathbf{m}}_i)$, we omit the semantic index i for brevity. The student predicts K candidate mask logits $\{M^k\}_{k=1}^K$, from which we select the candidate that best matches the pseudo-mask and supervise it using the binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ and the Dice loss $\mathcal{L}_{\text{dice}}$:

$$k^* = \arg\min_{k\in\{1,\dots,K\}} \mathcal{L}_{\text{dice}}\left(\sigma(M^k), \hat{\mathbf{m}}\right), \tag{4}$$

$$\mathcal{L}_{\text{pl}} = \mathcal{L}_{\text{bce}}\left(M^{k^*}, \hat{\mathbf{m}}\right) + \mathcal{L}_{\text{dice}}\left(\sigma(M^{k^*}), \hat{\mathbf{m}}\right), \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Semantic Prompt Consistency. For the same thermal image and semantic concept, the model should produce consistent predictions across semantically equivalent prompt expressions, which is important for robust open-vocabulary recognition. To enforce this consistency, we use the canonical prompt p_c to establish a semantic reference and randomly sample a prompt variant p_v from the corresponding prompt set. Then the semantic prompt consistency loss is defined as

$$M_v = f_{\theta_S}\big(\text{StrongAug}(\mathbf{x}^T), p_v\big), \qquad M_c = \text{sg}\left(f_{\theta_S}\big(\text{StrongAug}(\mathbf{x}^T), p_c\big)\right), \tag{6}$$

$$\mathcal{L}_{\text{pc}} = \mathcal{L}_{\text{dice}}\left(\sigma(M_v), \sigma(M_c)\right), \tag{7}$$

where f_{θ_S} denotes the student model, $\text{StrongAug}(\cdot)$ denotes the strong augmentation operation, and $\text{sg}(\cdot)$ stops gradient propagation.

EMA Teacher-Student Prediction Consistency. The pseudo-label supervision in Eq. 5 relies on fixed offline pseudo-masks and therefore cannot benefit from model improvements during adaptation. By contrast, the EMA teacher evolves with the student and can provide progressively refined online prediction targets. To exploit these dynamic targets, we introduce a teacher-student prediction consistency loss:

$$M_T = f_{\theta_T}\big(\text{WeakAug}(\mathbf{x}^T), p_c\big), \qquad M_S = f_{\theta_S}\big(\text{StrongAug}(\mathbf{x}^T), p_v\big), \tag{8}$$

$$\mathcal{L}_{\text{ts}} = \mathcal{L}_{\text{dice}}\left(\sigma(M_S), \mathcal{T}_{\text{weak}\to\text{strong}}\big(\text{sg}(\sigma(M_T))\big)\right). \tag{9}$$

where f_{θ_T} denotes the teacher model, $\text{WeakAug}(\cdot)$ denotes the weak augmentation operation, and $\mathcal{T}_{\text{weak}\to\text{strong}}$ warps the teacher prediction into the coordinate frame of the strongly augmented student view. In this way, the student benefits from the teacher's dynamically updated predictions and learns robust semantic responses under joint visual and linguistic perturbations.

Finally, the overall training objective for the second stage is defined as

$$\mathcal{L} = \lambda_{\text{pl}}\mathcal{L}_{\text{pl}} + \lambda_{\text{pc}}\mathcal{L}_{\text{pc}} + \lambda_{\text{ts}}\mathcal{L}_{\text{ts}}, \tag{10}$$

where λ_{pl}, λ_{pc}, and λ_{ts} control the contributions of the corresponding loss terms.

4 Experiments

4.1 Thermal-Only OVSS Benchmark and Evaluation Protocol

Benchmark. Most existing open-vocabulary semantic segmentation benchmarks focus primarily on RGB imagery. To evaluate thermal-only OVSS, we construct an evaluation benchmark suite based on FMB (Liu et al., 2023) and TIR-SS (Chen et al., 2022). FMB is a publicly available RGB–thermal dataset containing 1,500 spatially aligned image pairs with pixel-level annotations for 14 semantic categories, from which we retain only the thermal modality. TIR-SS contains 1,050 thermal images with pixel-level annotations for 8 urban-scene categories. After merging semantically equivalent categories across the two datasets, the resulting evaluation vocabulary contains 16 unique semantic concepts. To evaluate robustness to different textual expressions, we construct three prompts for each concept: a canonical category name, a synonymous expression, and a descriptive phrase. This results in 48 evaluation prompts, which are semantically disjoint from the 76 concepts represented by 228 prompts in our training prompt bank.

Evaluation Protocol. To the best of our knowledge, no existing method is specifically designed for thermal-only OVSS. We therefore compare ThermOV with representative RGB OVSS methods

under a unified evaluation protocol. **(1) ThermOV.** In Stage I, ThermOV is trained on COCO-Stuff (Caesar et al., 2018). After excluding prompt–mask pairs associated with the 16 evaluation concepts, 329,800 pairs covering 163 categories remain. In Stage II, we use the thermal modality of KAIST (Hwang et al., 2015). After removing images containing any evaluation concept, 7,493 thermal images are retained and treated as unlabeled, with pseudo-labels generated using our 76-concept training prompt bank. **(2) RGB OVSS.** Training-free methods are directly evaluated using their official checkpoints. For training-based methods, we report both direct-transfer and supervised-adaptation results, where supervised adaptation is performed on SODA (Li et al., 2021). SODA contains 2,168 real thermal images with pixel-level annotations for 20 foreground categories. After removing masks overlapping with the 16 evaluation concepts, 1,981 valid mask samples from 1,483 images over 11 categories remain for supervised adaptation.

4.2 Implementation Details

All experiments are conducted on two NVIDIA GeForce RTX 5090 GPUs. We train ThermOV using the AdamW optimizer with a learning rate of 1×10^{-5}. Stage I and Stage II are optimized for 800 and 1,600 steps, respectively. The input resolution is fixed at 1008×1008 in both stages. We use a per-GPU batch size of 1, resulting in a global batch size of 2, with each sample consisting of one prompt–mask pair. In Stage II, the loss weights $(\lambda_{pl}, \lambda_{ts}, \lambda_{pc})$ are set to $(0.95, 0.075, 0.005)$. Additional implementation details are provided in *Suppl.* Tables S3 and S7.

4.3 Comparative Results

Overall Performance. We compare ThermOV with representative RGB OVSS methods under a unified evaluation protocol. As shown in Table 1, ThermOV achieves 52.39 and 62.14 mIoU on FMB and TIR-SS, outperforming the original SAM3 by 5.97 and 9.30 mIoU, respectively. Although training-based methods benefit from additional pixel-level thermal supervision, ThermOV still surpasses the strongest SODA-adapted baselines by 3.81 and 0.54 mIoU, respectively. These results demonstrate effective thermal open-vocabulary adaptation without real thermal annotations.

Table 1: Comparison with representative OVSS methods on the thermal-only benchmark. Methods marked with † are adapted on pixel-annotated SODA images, while unmarked baselines use their official checkpoints. Best results are shown in bold.

Method	Backbone	FMB		TIR-SS	
		mIoU	mAcc	mIoU	mAcc
Training-Free OVSS					
ITA-CLIP (Aydın et al., 2025)	CLIP ViT-B	18.99	32.33	41.06	53.10
FreeCP (Chen et al., 2025)	CLIP ViT-B	18.07	29.66	39.52	52.90
Trident (Shi et al., 2025)	CLIP ViT-B + DINO ViT-B	23.75	35.73	52.68	63.70
NACLIP (Hajimiri et al., 2025)	CLIP ViT-B	22.02	36.16	44.63	57.10
CASS (Kim et al., 2025)	CLIP ViT-B + DINO ViT-B	24.05	39.08	47.31	60.05
SAM3 (Carion et al., 2026)	SAM3-ViT	46.42	51.61	52.84	57.82
Training-Based OVSS					
OVSeg (Liang et al., 2023)	Swin-B + CLIP ViT-L	32.84	44.70	51.76	68.50
OVSeg†		35.79	47.79	54.35	71.78
EBSeg (Shan et al., 2024)	CLIP ViT-B + SAM ViT-B	34.92	48.92	59.11	74.70
EBSeg†		39.44	53.32	62.50	**76.10**
CAT-Seg (Cho et al., 2024)	CLIP ViT-B	36.14	46.46	57.61	75.47
CAT-Seg†		35.85	46.28	57.12	75.34
Mask-Adapter (Li et al., 2025)	OpenCLIP ConvNeXt-L	43.81	54.24	60.78	74.99
Mask-Adapter†		44.94	55.37	61.60	75.60
ThermOV (ours)	SAM3-ViT	**52.39**	**58.20**	**62.14**	68.80

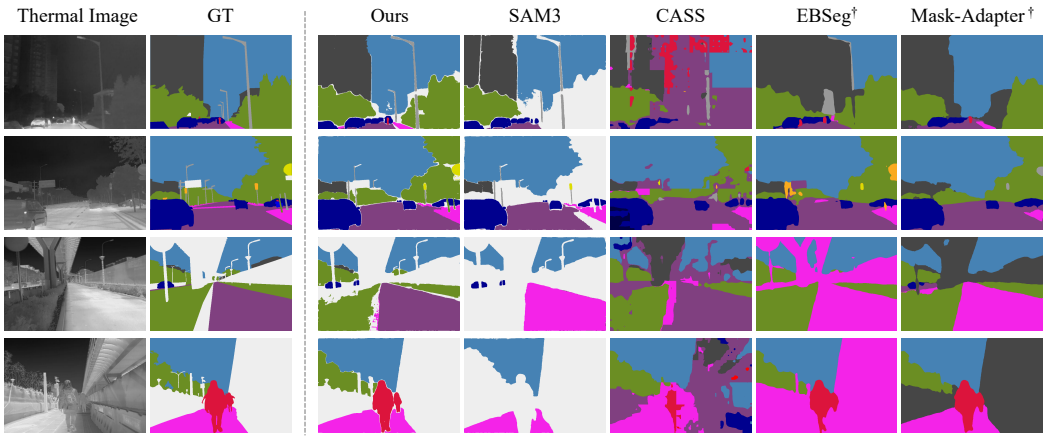

Figure 4: Qualitative comparison on FMB and TIR-SS. ThermOV produces more accurate predictions and object boundaries. † denotes adaptation on pixel-annotated SODA images.

Qualitative Comparison. Figure 4 presents qualitative comparisons on FMB and TIR-SS. Compared with representative training-free methods (SAM3 and CASS) and SODA-adapted training-based methods (EBSeg† and Mask-Adapter†), ThermOV produces more accurate target regions and boundaries in challenging thermal scenes, with less semantic confusion and fewer spurious responses.

Robustness across Prompt Variants. We evaluate robustness to textual variations using canonical, synonym, and descriptive prompts. As shown in Table 2, SAM3 and CASS exhibit larger performance variations across prompt types, indicating higher sensitivity to prompt wording, whereas the SODA-adapted EBSeg† and Mask-Adapter† show smaller variations. Although Mask-Adapter† achieves the lowest Prompt Std., ThermOV obtains the highest mean mIoU across all three prompt types while maintaining relatively small performance variations.

Table 2: Robustness to prompt variations on FMB and TIR-SS. Mean mIoU averages the mIoU values on the two datasets for each prompt type, while Prompt Std. measures its standard deviation across the three prompt types. Methods marked with † are adapted on pixel-annotated SODA images. Best results are shown in bold.

Method	Prompt	FMB		TIR-SS		Mean mIoU↑	Prompt Std.↓
		mIoU	mAcc	mIoU	mAcc		
CASS	Canonical	24.02	39.06	47.31	60.05	35.66	1.946
	Synonym	21.54	34.09	40.28	59.55	30.91	
	Descriptive	26.64	39.81	40.50	64.34	33.57	
SAM3	Canonical	46.42	51.61	52.84	57.82	49.63	6.453
	Synonym	31.30	33.69	36.47	39.73	33.88	
	Descriptive	39.99	42.82	45.92	49.86	42.95	
EBSeg†	Canonical	39.44	53.32	62.50	76.10	50.97	1.551
	Synonym	35.53	48.70	59.13	72.39	47.33	
	Descriptive	36.65	48.99	63.55	75.40	50.10	
Mask-Adapter†	Canonical	45.02	55.42	61.90	75.77	53.46	**0.558**
	Synonym	45.09	53.71	59.25	73.71	52.17	
	Descriptive	43.72	53.25	62.71	74.58	53.22	
ThermOV	Canonical	**52.39**	**58.20**	**62.14**	68.81	**57.26**	1.241
	Synonym	**51.52**	**56.37**	**57.09**	62.56	**54.31**	
	Descriptive	**52.09**	**56.91**	**60.71**	66.52	**56.40**	

4.4 Ablation Study

Component Ablation. We ablate Stage-I pseudo-thermal adaptation and three Stage-II objectives: pseudo-label supervision, teacher–student prediction consistency, and semantic prompt consistency. As shown in Table 3, Stage-I adaptation improves SAM3 by 0.72 and 2.55 mIoU on FMB and TIR-SS, respectively, providing a better initialization for Stage II. Pseudo-label supervision yields substantial gains, and teacher–student consistency further improves performance. Prompt consistency is not consistently beneficial when added alone, but complements teacher–student consistency and leads to the best overall performance.

Qualitative Analysis of Stage-I Adaptation. Figure 5 compares the pseudo-labels generated by SAM3 and SAM3-PTA on representative thermal images, with several samples manually annotated as GT for reference. Compared with SAM3, SAM3-PTA produces more complete target regions with fewer missing predictions.

Table 3: Component ablation of ThermOV on FMB and TIR-SS.

Stage-I PTA	Stage-II Training			FMB		TIR-SS	
	PL Sup.	T–S Cons.	Prompt Cons.	mIoU	mAcc	mIoU	mAcc
×	×	×	×	46.42	51.61	52.84	57.82
✓	×	×	×	47.14	52.22	55.39	60.90
×	✓	✓	✓	48.04	53.72	53.96	59.49
✓	✓	×	×	49.84	55.22	60.09	66.40
✓	✓	✓	×	51.13	57.09	61.60	68.12
✓	✓	×	✓	49.79	54.89	57.75	63.25
✓	✓	✓	✓	**52.39**	**58.20**	**62.14**	**68.80**

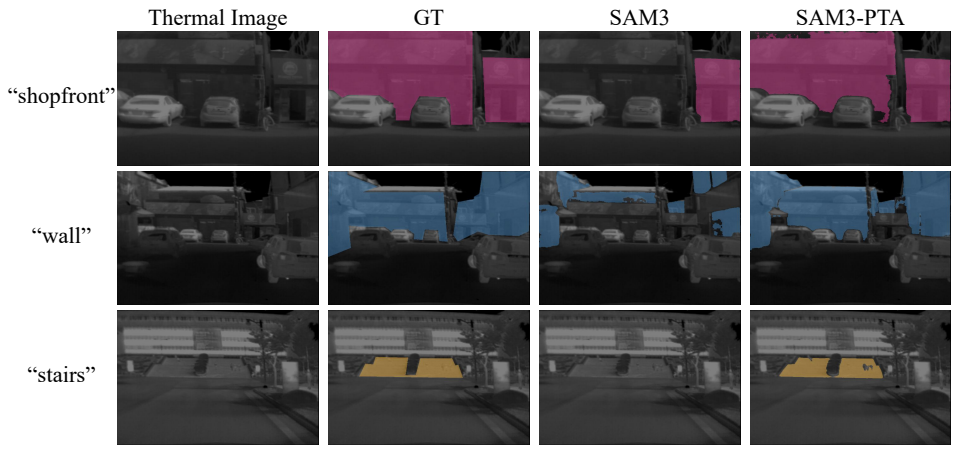

Figure 5: Qualitative comparison of pseudo-labels generated by SAM3 and SAM3-PTA on representative thermal images. GT masks are manually annotated for reference.

5 Conclusion

In this paper, we present ThermOV, a thermal-only open-vocabulary semantic segmentation method based on unsupervised cross-domain adaptation. ThermOV employs a two-stage strategy comprising pseudo-thermal initialization and unlabeled real-thermal adaptation, with pseudo-label supervision, semantic prompt consistency, and teacher–student consistency. Experiments on FMB and TIR-SS demonstrate that ThermOV substantially improves thermal open-vocabulary segmentation and outperforms strong RGB OVSS baselines, including those adapted using pixel-level thermal annotations. These results demonstrate the potential of unsupervised cross-domain adaptation for extending vision foundation models to thermal imagery.

AI USAGE STATEMENT

During the preparation of this manuscript, we used generative AI tools for language refinement, grammar checking, and proofreading assistance. All scientific contributions, including the proposed methodology, experimental design, data analysis, and conclusions, were developed and verified by the authors. The authors take full responsibility for the final content of this manuscript.

REFERENCES

M. Arda Aydın, Efe Mert Çırpar, Elvin Abdinli, Gozde Unal, and Yusuf H. Sahin. ITACLIP: Boosting training-free semantic segmentation with image, text, and architectural enhancements. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4181–4191, 2025.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, 2026.

Junzhang Chen, Zhong Liu, Fan Yang, and Xiangzhi Bai. Light transport induced domain adaptation for semantic segmentation in thermal infrared urban scenes. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):23194–23211, 2022.

Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 23124–23134, 2025.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Lu Gan, Connor Lee, and Soon-Jo Chung. Unsupervised RGB-to-thermal domain adaptation via multi-domain attention network. In *IEEE International Conference on Robotics and Automation*, pp. 6014–6020, 2023.

Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5061–5071, 2025.

Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9924–9935, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1037–1045, 2015.

Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15033–15042, 2025.

Chenglong Li, Wei Xia, Yan Yan, Bin Luo, and Jin Tang. Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):3069–3082, 2021.

Yongkang Li, Tianheng Cheng, Bin Feng, Wenyu Liu, and Xinggang Wang. Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14998–15008, 2025.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 8115–8124, 2023.

Silvio Mazzucco, Carl Persson, Mattia Segu, Pier Luigi Dovesi, Federico Tombari, Luc Van Gool, and Matteo Poggi. Lost in translation? vocabulary alignment for source-free adaptation in open-vocabulary semantic segmentation. In *British Machine Vision Conference*, 2025.

Tayeba Qazi, Ayush Maheshwari, Prerana Mukherjee, and Brejesh Lall. T-CLIP: Enabling thermal perception for contrastive language-image pretraining, 2026.

Reza Qorbani, Gianluca Villani, Theodoros Panagiotakopoulos, Marc Botet Colomer, Linus Härenstam-Nielsen, Mattia Segu, Pier Luigi Dovesi, Jussi Karlgren, Daniel Cremers, Federico Tombari, and Matteo Poggi. Semantic library adaptation: LoRA retrieval and fusion for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9804–9815, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Yasiru Ranasinghe, Elim Schenck, Florence Yellin, Shuowen Hu, Christopher Funk, and Vishal M. Patel. Thermal-det: Language-guided cross-modal distillation for open-vocabulary thermal object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34628–34637, 2026.

Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28412–28421, 2024.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *IEEE/CVF International Conference on Computer Vision*, pp. 23487–23497, 2025.

Aniruddh Sikdar, Aditya Gandhamal, and Suresh Sundaram. Aethervision-bench: An open-vocabulary RGB-infrared benchmark for multi-angle segmentation across aerial and ground perspectives, 2025. Presented at the CVPR 2025 Workshop on Foundation Models Meet Embodied Agents (non-archival).

Young-Chae Son, Dae-Kwan Ko, Yoon-Ji Choi, and Soo-Chul Lim. Thermoact: Thermal-aware vision-language-action models for robotic perception and decision-making. *IEEE Robotics and Automation Letters*, 11(5):6106–6113, 2026.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1195–1204, 2017.

Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1379–1389, 2021.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Meng Yu, Yufeng Yue, Luojie Yang, Xunjie He, Yi Yang, and Mengyin Fu. Open-RGBT: Open-vocabulary RGB-T zero-shot semantic segmentation in open-world environments. In *IEEE International Conference on Robotics and Automation*, pp. 10729–10735, 2025.

Tao Zhang, Jinyong Wen, Zhen Chen, Kun Ding, Shiming Xiang, and Chunhong Pan. UNIP: Rethinking pre-trained attention patterns for infrared semantic segmentation. In *International Conference on Learning Representations*, 2025.

Guoqiang Zhao, Junjie Huang, Xiaoyun Yan, Zhaojing Wang, Junwei Tang, Yangjun Ou, Xinrong Hu, and Tao Peng. Open-vocabulary RGB-thermal semantic segmentation. In *European Conference on Computer Vision*, pp. 304–320, 2024.

Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11350–11359, 2023.

Xue-Feng Zhu, Tianyang Xu, Yifan Pan, Jinjie Gu, Xi Li, Jiwen Lu, Xiaojun Wu, and Josef Kittler. Collaborating vision, depth, and thermal signals for multi-modal tracking: Dataset and algorithm. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

A APPENDIX

A.1 STAGE-I: PSEUDO-THERMAL DATA GENERATION

To construct thermal-style supervision from labeled RGB data, we generate pseudo-thermal images through a series of transformations, including luminance conversion, category-aware thermal prior injection, and thermal appearance perturbation. The generated pseudo-thermal samples preserve the original semantic annotations while introducing thermal-specific appearance characteristics for visual adaptation.

A.1.1 LUMINANCE CONVERSION AND THERMAL PRIOR INJECTION

In Stage-I pseudo-thermal generation, RGB images are first transformed into single-channel intensity representations and further enhanced with category-aware thermal priors. To simulate the different thermal responses of semantic regions, we manually assign intensity offsets according to the typical thermal characteristics of different categories.

Specifically, categories corresponding to heat-emitting objects are assigned positive offsets, while cold background regions are assigned negative offsets. Structural regions with relatively stable thermal responses receive small positive offsets. The unmatched categories are assigned zero offsets. The detailed thermal prior values used in our experiments are summarized in Table S1.

Table S1: Category-specific thermal prior values used for pseudo-thermal generation.

Category Group	Category	Thermal Prior Δ
Heat-emitting objects	person	+0.34
	rider	+0.30
	car	+0.24
	truck	+0.26
	bus	+0.27
Urban structures	building	+0.035
	wall	+0.035
	road	+0.040
	sidewalk	+0.030
Cold background	sky	-0.16
	cloud	-0.12
	vegetation	-0.08
	tree	-0.08
Others	unmatched categories	0.00

A.1.2 THERMAL APPEARANCE PERTURBATION AND POST-PROCESSING

After thermal prior injection, we further apply image-level appearance perturbations to simulate variations in thermal imaging conditions, including intensity response differences, contrast changes, brightness shifts, and sensor noise. Finally, lightweight post-processing operations are applied to enhance the diversity of generated thermal appearances.

The detailed perturbation and post-processing configurations are listed in Table S2.

A.2 STAGE-I: PARAMETER-EFFICIENT VISUAL ADAPTATION

Based on the constructed pseudo-thermal supervision, we adapt the visual representation of SAM3 toward thermal characteristics in a parameter-efficient manner. During training, the original SAM3 parameters are frozen, including the text encoder, visual encoder weights, and segmentation head. LoRA adapters are introduced into the visual encoder to update only a small number of parameters while preserving the pretrained open-vocabulary semantic space.

Table S2: Pseudo-thermal appearance perturbation and post-processing configurations.

Operation	Probability	Parameter
Gamma correction	1.0	$\gamma \sim U(0.78, 1.25)$
Contrast adjustment	1.0	factor $\sim U(1.05, 1.38)$
Brightness shift	1.0	offset $\sim U(-0.04, 0.07)$
Gaussian noise	1.0	$\sigma \sim U(0.012, 0.040)$
Autocontrast	1.0	cutoff=0.5
Gaussian blur	0.1	radius $\sim U(0.25, 0.8)$
Sharpening	0.5	radius=1, percent=120

Specifically, LoRA adapters are inserted into the last eight transformer blocks of the visual encoder and applied to the attention projection layers. The detailed Stage-I training configurations are summarized in Table S3.

Table S3: Stage-I visual adaptation settings.

Configuration	Setting
Base model	SAM3
Trainable module	LoRA adapters
Adapted layers	Last 12 visual transformer blocks
LoRA modules	Attention projection layers
LoRA rank	8
LoRA alpha	16
LoRA dropout	0.0
Optimizer	AdamW
Weight decay	0.01
Learning rate	1×10^{-5}
Input resolution	1008×1008
Training steps	800
Batch size	1 prompt-mask sample/step
Gradient clipping	Maximum norm 1.0

A.3 Stage-II: Unlabeled Thermal Adaptation

After obtaining the thermal-adapted initialization model from Stage-I, we further perform adaptation on real unlabeled thermal images. During this stage, ThermOV leverages semantic prompts to construct prompt-conditioned pseudo-labels and introduces teacher-student consistency learning to exploit unlabeled thermal data.

A.3.1 Prompt Bank Construction

We constructs a semantic prompt bank to generate prompt-conditioned pseudo labels from unlabeled thermal images and support semantic prompt consistency learning during thermal adaptation. The prompt bank contains 76 manually curated semantic concepts, covering common visual entities in urban environments, road scenes, infrastructures, and daily objects. For each semantic concept, multiple natural language expressions are to describe the same visual concept from different textual perspectives.

Based on this prompt bank, SAM3-PTA generates 16,996 prompt-mask pseudo-label pairs from unlabeled thermal images during the real thermal adaptation stage. Representative examples of the semantic prompt bank are shown in Table S4. To further evaluate the robustness of ThermOV under different textual formulations, we construct three types of evaluation prompts: canonical prompts, synonym prompts, and descriptive prompts. Canonical prompts directly use the original category names from the datasets, synonym prompts replace them with semantically equivalent expressions,

Table S4: Representative examples of the semantic prompt bank.

Concept	Prompt variants
road	road; street; paved road surface
lane line	lane line; painted lane line; white lane divider
curb	curb; curb edge; raised curb
building	building; building structure; urban facade
pole	pole; utility pole; roadside pole
vehicle	vehicle; automobile; road vehicle
cable	cable; electric cable; coiled cable
billboard	billboard; advertising board; display board
awning	awning; shop awning; fabric canopy
insulation	insulation; insulation material; thermal insulation

and descriptive prompts represent the same concepts using natural language descriptions. The evaluation prompt examples are presented in Table S5.

Table S5: Examples of three prompt types used for robustness evaluation.

Dataset	Canonical	Synonym	Descriptive
FMB	road	street	drivable road surface
FMB	person	human	standing walking person
FMB	car	automobile	passenger car
TIR-SS	vehicle	automobile	vehicle on the road
TIR-SS	vegetation	greenery	trees and vegetation
TIR-SS	building	building structure	urban building facade

A.3.2 Teacher–Student View Construction

During real thermal adaptation, we adopt different image views for teacher and student networks to enforce prediction consistency. The teacher receives weakly augmented thermal images to generate stable pseudo labels, while the student is optimized using mildly perturbed images.

The detailed augmentation configurations are summarized in Table S6.

Table S6: Teacher-student augmentation configurations during Stage-II adaptation.

Operation	Teacher View	Student View
Resize	1008×1008	1008×1008
Horizontal flip	Disabled	$p = 0.5$
Contrast adjustment	None	factor $\sim U(0.97, 1.03)$, $p = 0.9$
Brightness adjustment	None	factor $\sim U(0.97, 1.03)$, $p = 0.9$

A.3.3 EMA Teacher and Multi-Constraint Optimization

The teacher network is updated using exponential moving average (EMA) of the student parameters. During Stage-II adaptation, the model is optimized with three complementary objectives, including pseudo-label supervision, teacher-student consistency constraint, and semantic prompt consistency constraint.

The detailed optimization configurations are summarized in Table S7.

Table S7: Stage-II adaptation settings.

Configuration	Setting
Initialization model	SAM3-PTA
Input resolution	1008×1008
Optimizer	AdamW
Stage-II learning rate (LoRA)	1×10^{-5}
Prediction head learning rate	5×10^{-5}
Training steps	1600
EMA decay	0.999
EMA warm-up steps	300
Pseudo-label loss weight λ_{pl}	0.95
Teacher-student consistency weight λ_{ts}	0.075
Prompt consistency weight λ_{pc}	0.005

A.4 Loss Weight Sensitivity

To examine the influence of the Stage-II loss weights, we vary one coefficient at a time around the default setting while keeping all other training configurations fixed. As shown in Table S8, ThermOV is relatively robust to moderate changes in λ_{ts}, whereas its performance is more sensitive to λ_{pl} and λ_{pc}. The default setting $(\lambda_{pl}, \lambda_{ts}, \lambda_{pc}) = (0.95, 0.075, 0.005)$ achieves the best overall performance on both FMB and TIR-SS.

Table S8: Sensitivity analysis of the loss weights in ThermOV.

λ_{pl}	λ_{ts}	λ_{pc}	FMB mIoU	FMB mAcc	TIR-SS mIoU	TIR-SS mAcc
0.900	0.075	0.005	43.095	47.336	49.097	53.805
0.950	**0.075**	**0.005**	**52.387**	**58.200**	**62.135**	**68.807**
1.000	0.075	0.005	47.425	52.613	55.313	61.367
0.950	0.050	0.005	50.553	56.248	60.765	67.058
0.950	0.100	0.005	51.357	57.636	61.727	68.532
0.950	0.075	0.0025	50.209	55.781	59.518	65.355
0.950	0.075	0.010	50.036	55.985	59.539	66.168